null

Data Analytics and Visualization

# of SP500 with Macroeconomics

Midterm Essay

*Presented by **Nguyen Quang Huy** - **Nguyen Gia Nhat Khanh***

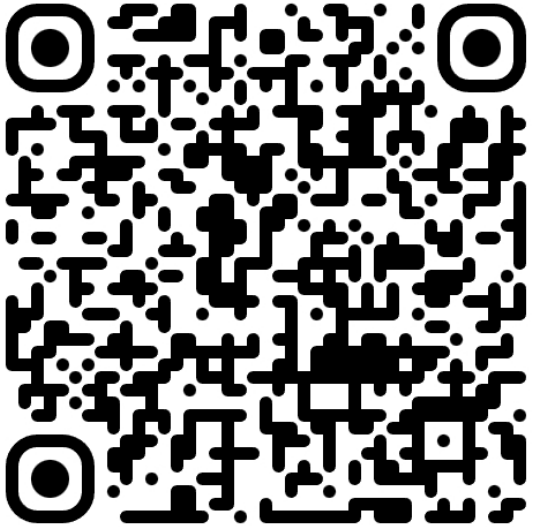
Faculty of Information Technology, Ton Duc Thang University

{nguyenquanghuy.st, 523h0149}@student.tdtu.edu.vn

QR Code

Presentation slide

Implementation



Github Resources

Outline

1. Introduction
2. Exploratory Data Analysis (EDA)
3. Probability Distribution Analysis
4. Hypothesis Testing
5. Correlation Analysis
6. Multiple Linear Regression
7. Conclusion

1. Introduction

- **Objective**: Analyze the dataset: EDA → distribution checks → hypothesis testing → correlation analysis → multiple linear regression.
- **Time-series dataset**:
  - Daily S&P 500 Index (close, open, high, low, volume) - Yahoo Finance.
  - Daily Macroeconomics, asset proxies (CPI, GLD, IEF, VIXY, UUP, USO) - Alpha Vantage.
  - Time range: Jan 4, 2011 to Dec12, 2023.
- **Research question:**

2. Exploratory Data Analysis (EDA)

Daily SP500 Index

- 24367 entries, 1927-12-30 to 2024-12-31.

Date	Open	High	Low	Close	Volume
2024-12-24	5984.629883	6040.100098	5981.439941	6040.040039	1757720000
2024-12-26	6024.970215	6049.750000	6007.370117	6037.589844	2904530000
2024-12-27	6006.169922	6006.169922	5932.950195	5970.839844	3159610000

2. Exploratory Data Analysis (EDA)

- **Daily Macroeconomics Index & Asset Proxies**

Proxies	Description
Consumer Price Index (**CPI**)	Measures inflation by tracking changes in consumer prices for goods and services.
SPDR Gold Shares (**GLD**)	Represents the price of gold through the SPDR Gold Shares ETF.
iShares 7–10 Year Treasury Bond (**IEF**)	Tracks U.S. Treasury bonds with 7–10-year maturities, reflecting long-term interest rates.
United States Oil Fund (**USO**)	Follows crude oil prices via the United States Oil Fund ETF.
Invesco DB U.S. Dollar Index Bullish Fund (**UUP**)	Reflects the strength of the U.S. dollar against major currencies through the Dollar Index ETF.
ProShares VIX Short-Term Futures (**VIXY**)	Represents market volatility based on short-term VIX futures, often called the “fear index.”

2. Exploratory Data Analysis (EDA)

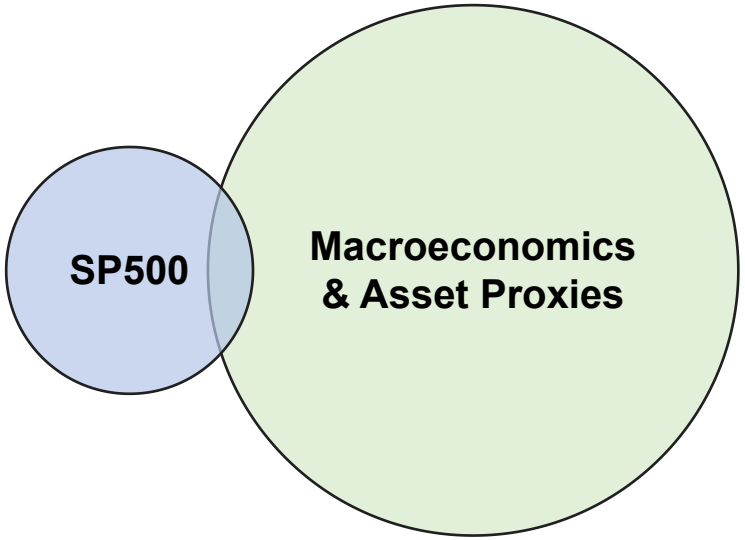
Daily Macroeconomics Index & Asset Proxies

- 6000 entries, 2000-24-30 to 2024-12-31.

Date	Open	High	Low	Close	Volume
2024-12-24	5984.629883	6040.100098	5981.439941	6040.040039	1757720000
2024-12-26	6024.970215	6049.750000	6007.370117	6037.589844	2904530000
2024-12-27	6006.169922	6006.169922	5932.950195	5970.839844	3159610000

2. Exploratory Data Analysis (EDA)

Dataset Merging



- 24367 entries, 11 features.
- Duplicate rows: 3856.
- Zero value: 5497 (Volume).
- Null value:

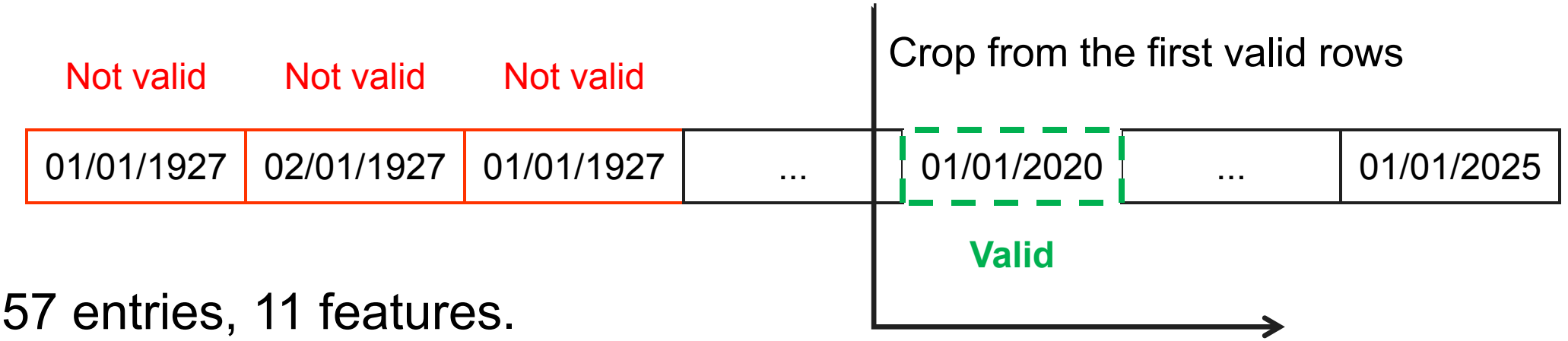
CPI: 20811 **USO**: 19653

GLD: 19304 **UUP**: 19876

IEF: 18720 **VIXY**: 20846

2. Exploratory Data Analysis (EDA)

Data Preprocessing



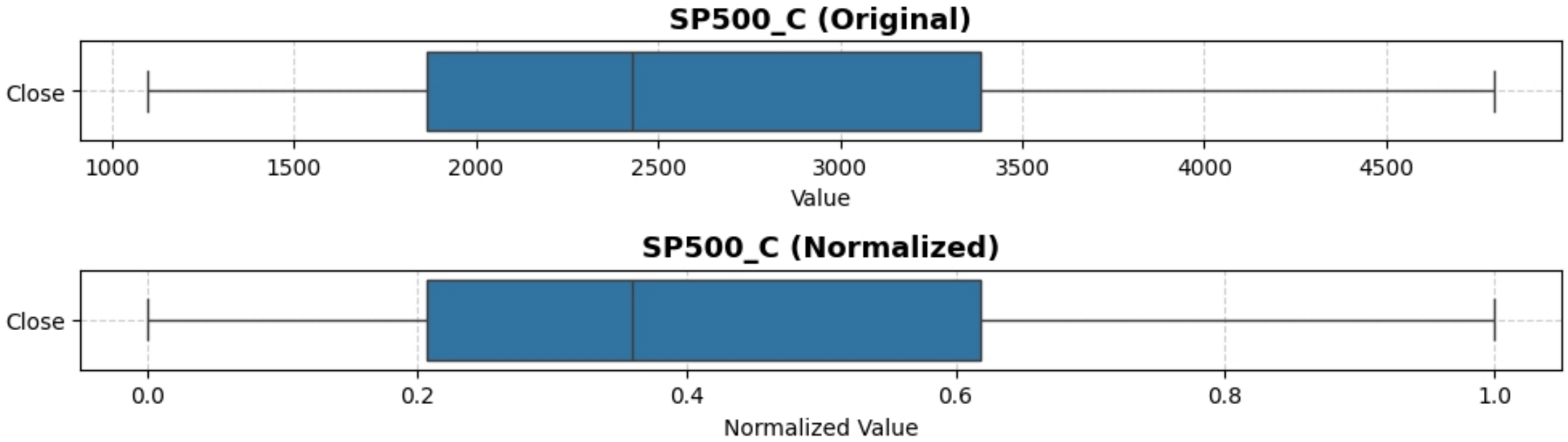
- 3257 entries, 11 features.
- Duplicate rows: 0.
- Zero value: 0.
- Null value: 0.

2. Exploratory Data Analysis (EDA)

Visualization (CandleStick) SP500 Index

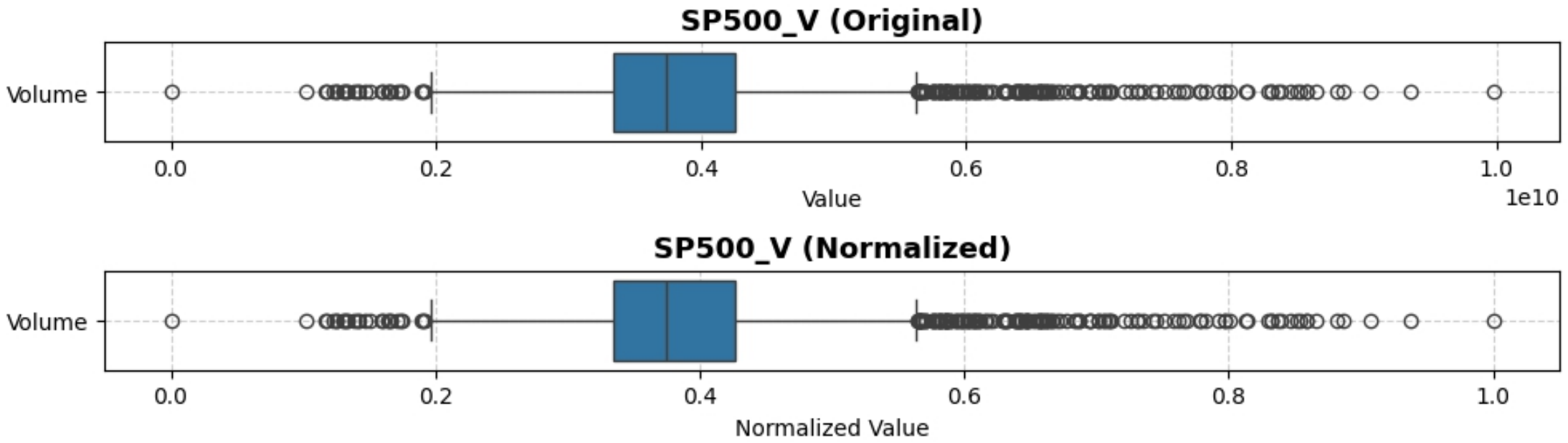
2. Exploratory Data Analysis (EDA)

Visualization (Boxplot) Close Index



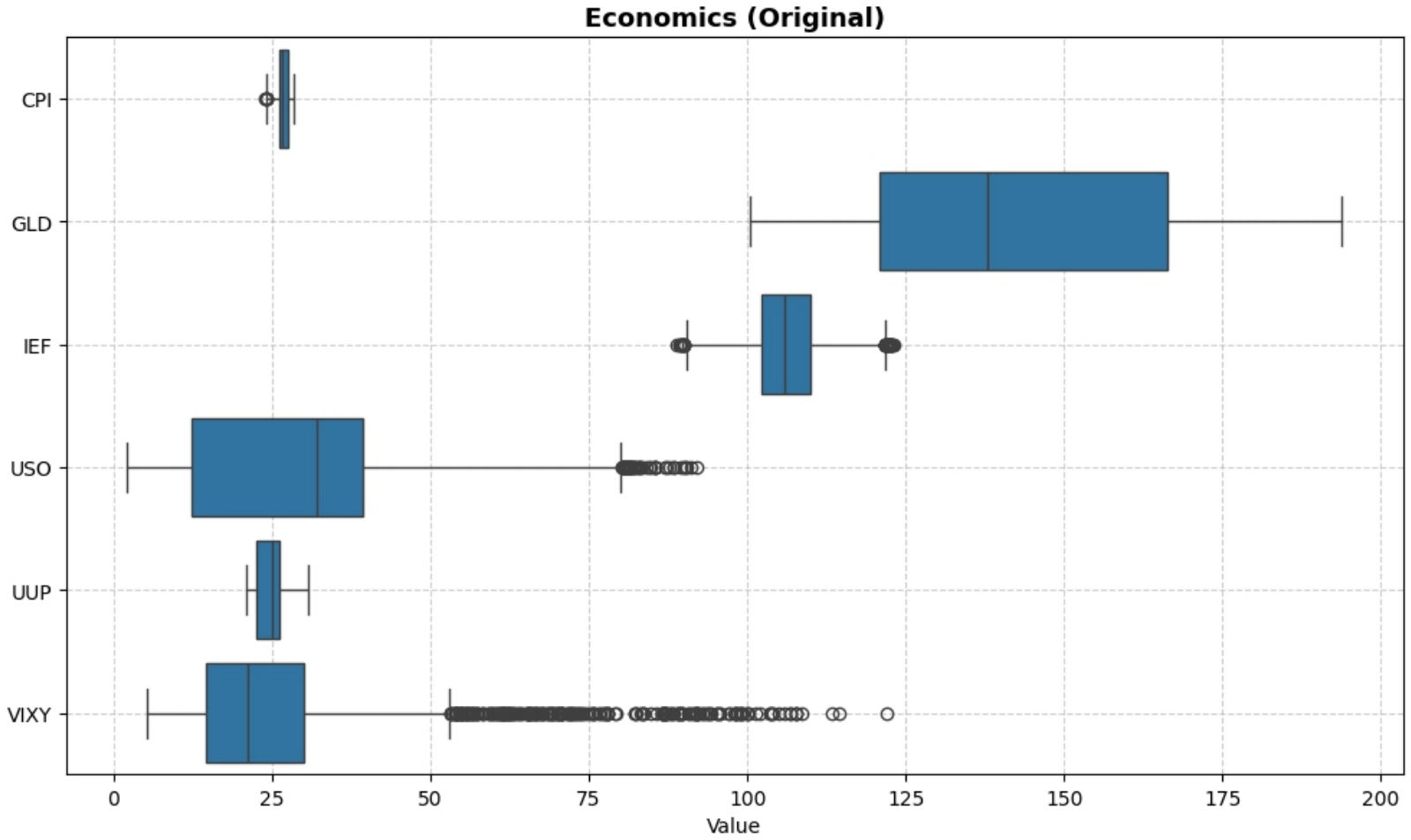
2. Exploratory Data Analysis (EDA)

Visualization (Boxplot) Trading Volume



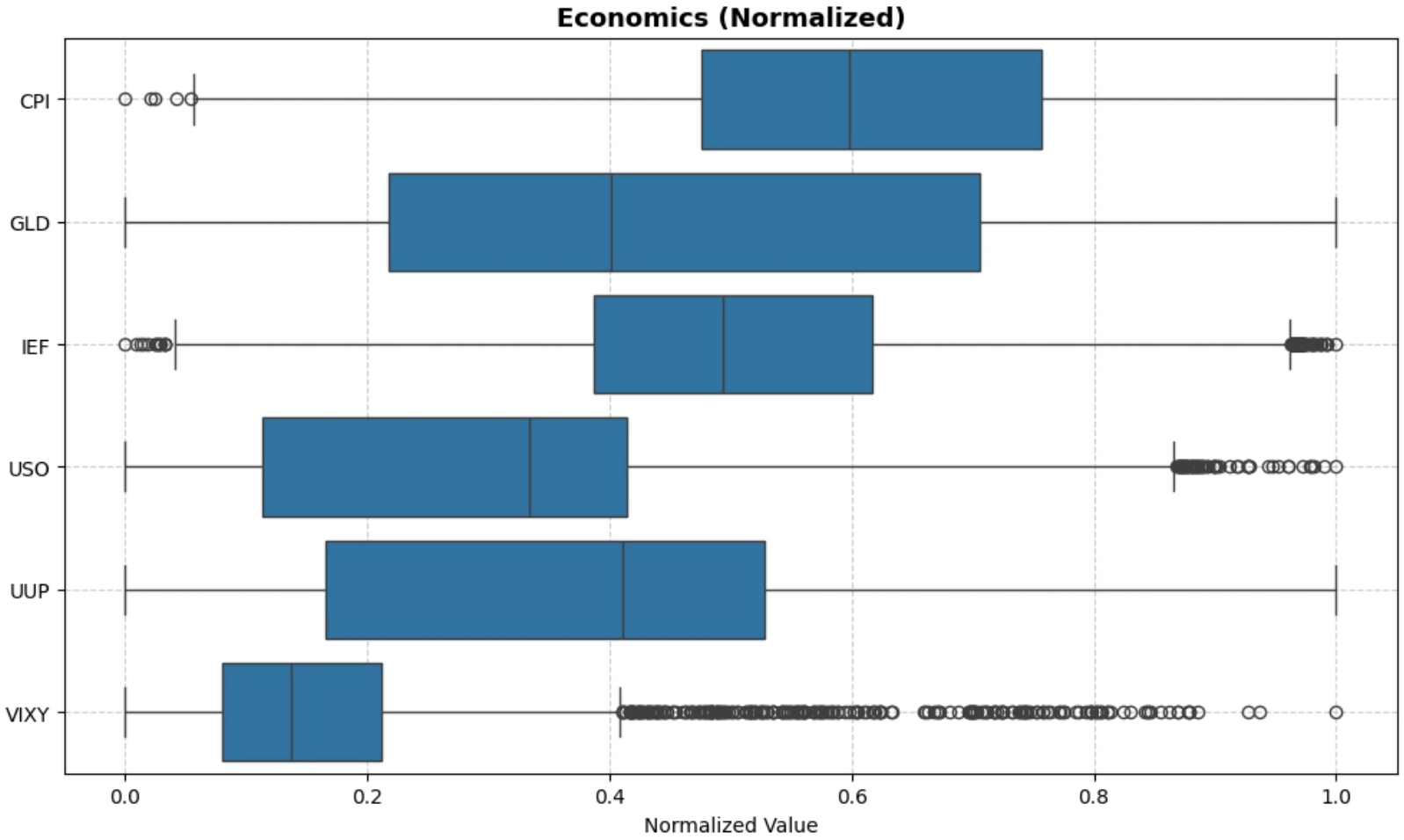
2. Exploratory Data Analysis (EDA)

Visualization (Boxplot) Macroeconomics



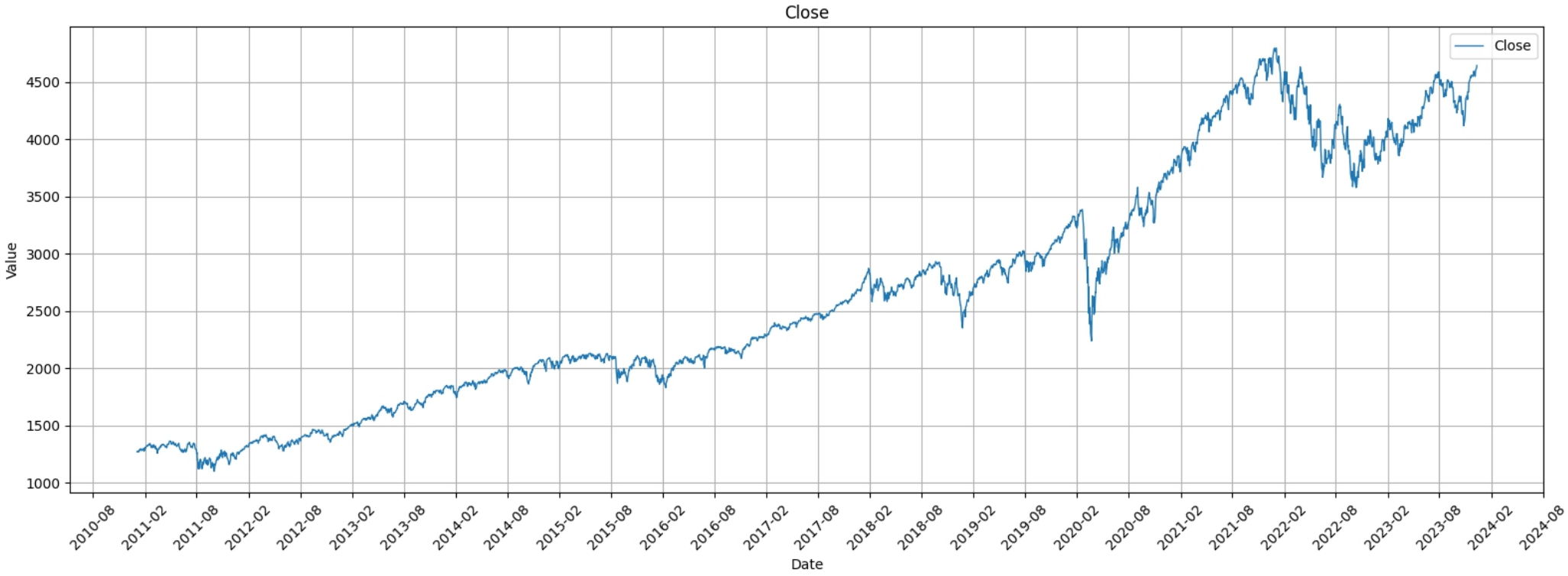
2. Exploratory Data Analysis (EDA)

Visualization (Boxplot) Macroeconomics



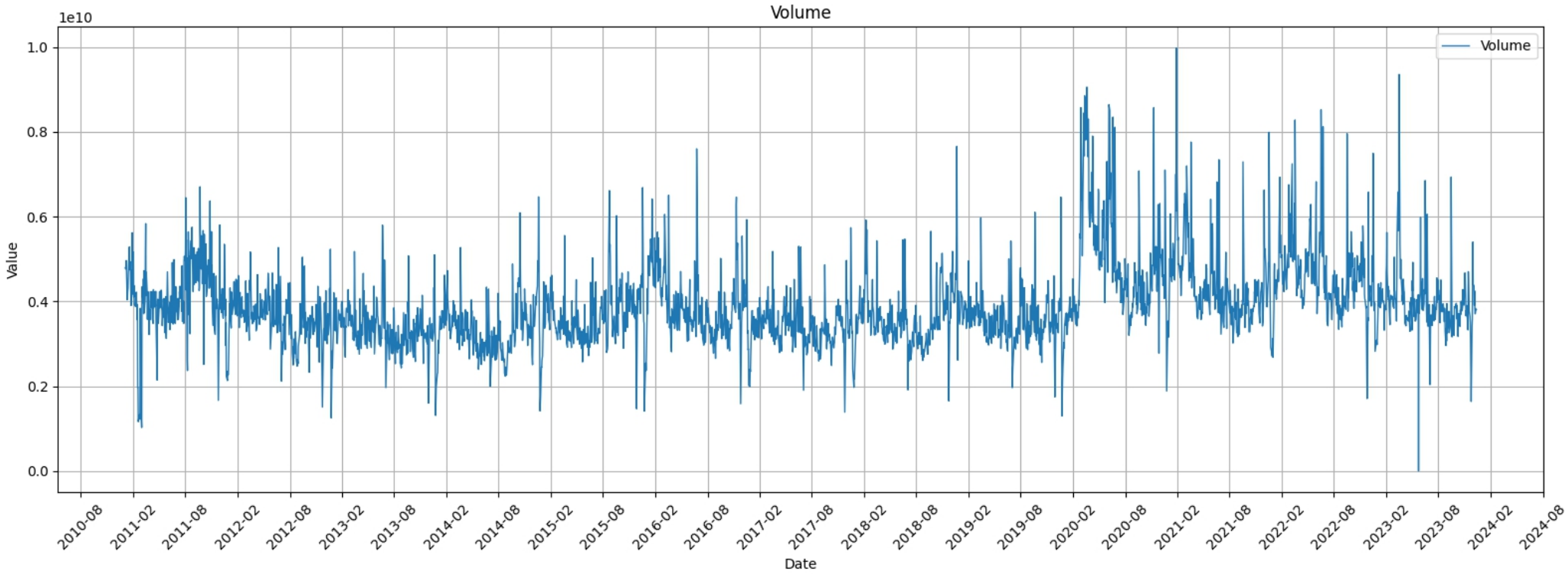
2. Exploratory Data Analysis (EDA)

Visualization (Line Chart) Close Index



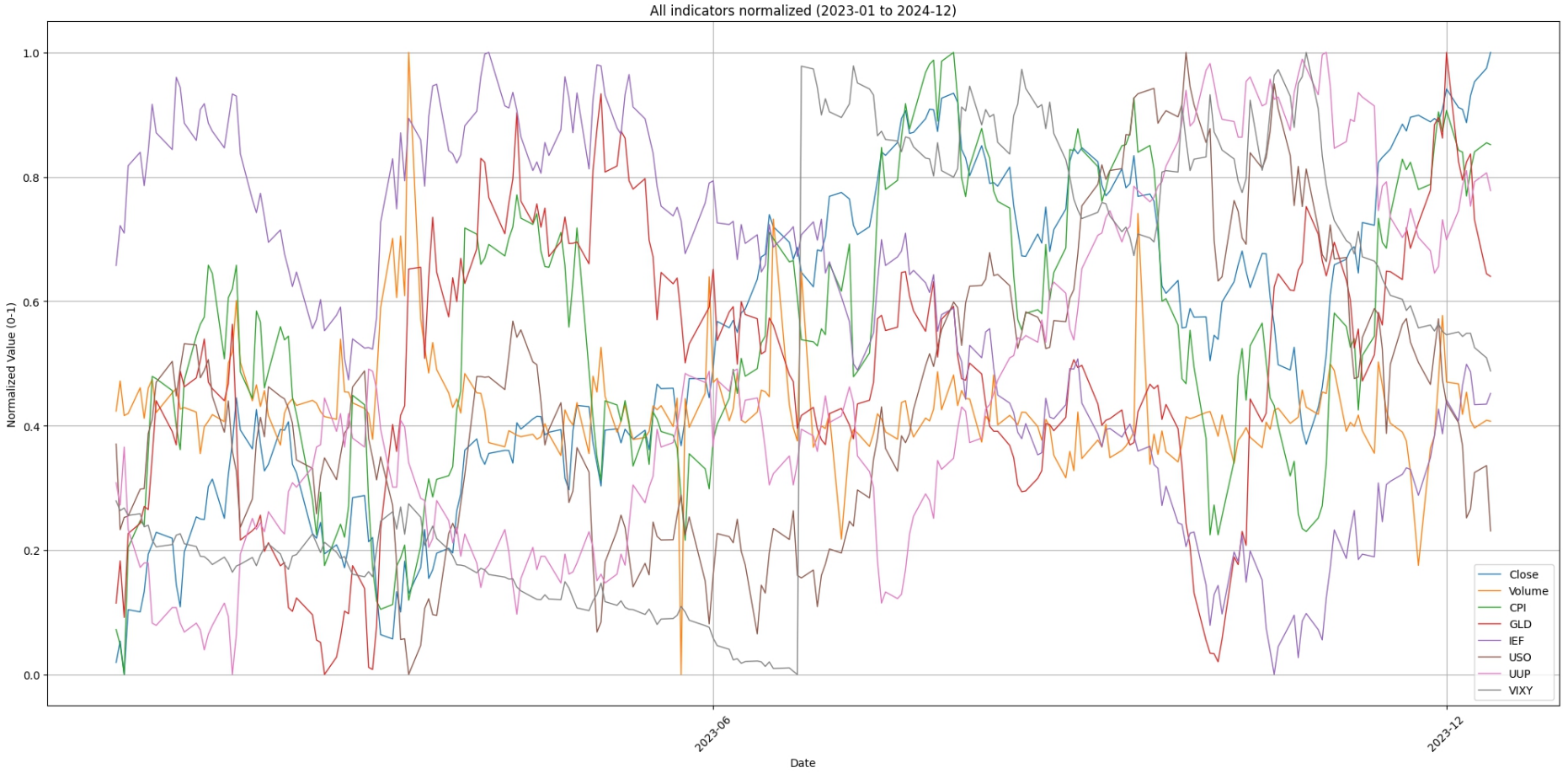
2. Exploratory Data Analysis (EDA)

Visualization (Line Chart) Close Index



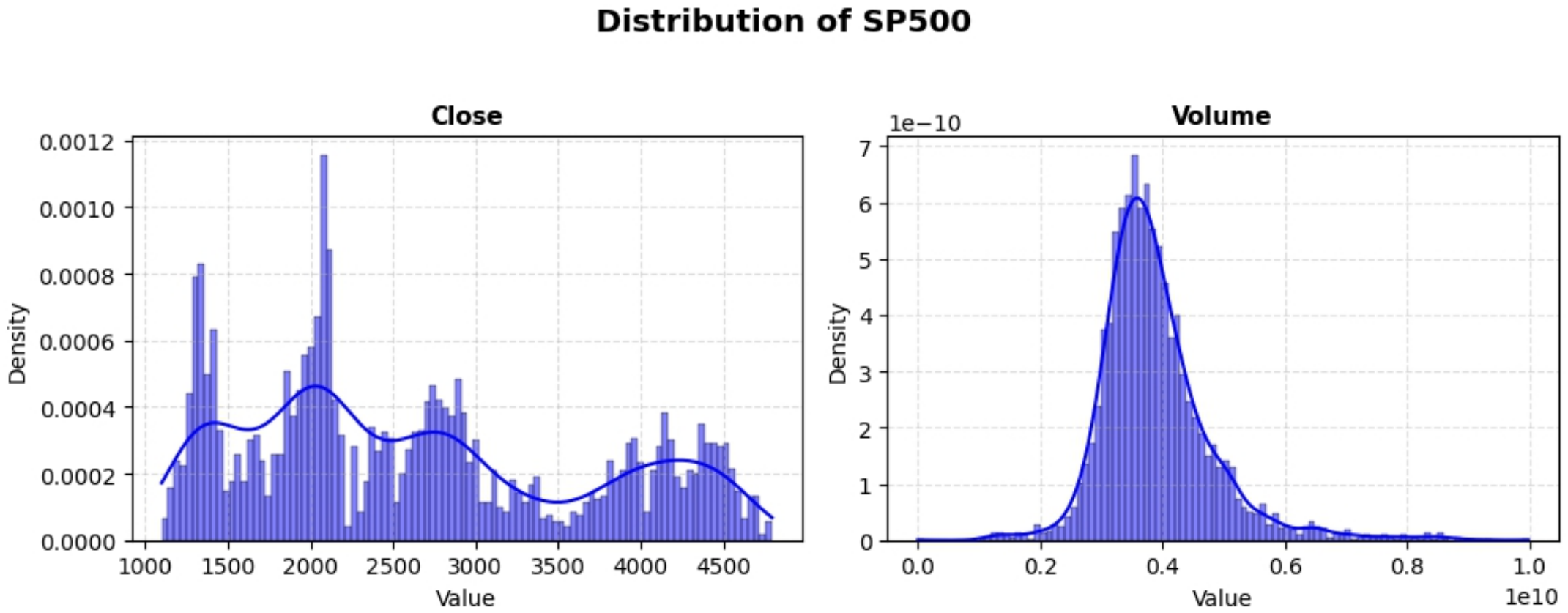
2. Exploratory Data Analysis (EDA)

Visualization (Line Chart) Close Index



2. Exploratory Data Analysis (EDA)

Visualization (Histogram) Close Index & Trading Volume



2. Exploratory Data Analysis (EDA)

Anomaly Detection

1. Z-score: Identifies outliers by measuring how many standard deviations a point is from the mean.

2. IQR: Detects outliers as values falling below Q1 − 1.5×IQR or above Q3 + 1.5×IQR.

3. Rolling Z-score: Applies Z-score on a moving window to find outliers in time series data.

4. Seasonal Decomposition Residuals: Flags outliers as large deviations in the residual component after decomposing trend and seasonality.

5. Isolation Forest: Uses an ensemble of random trees to isolate anomalies that require fewer splits to separate.

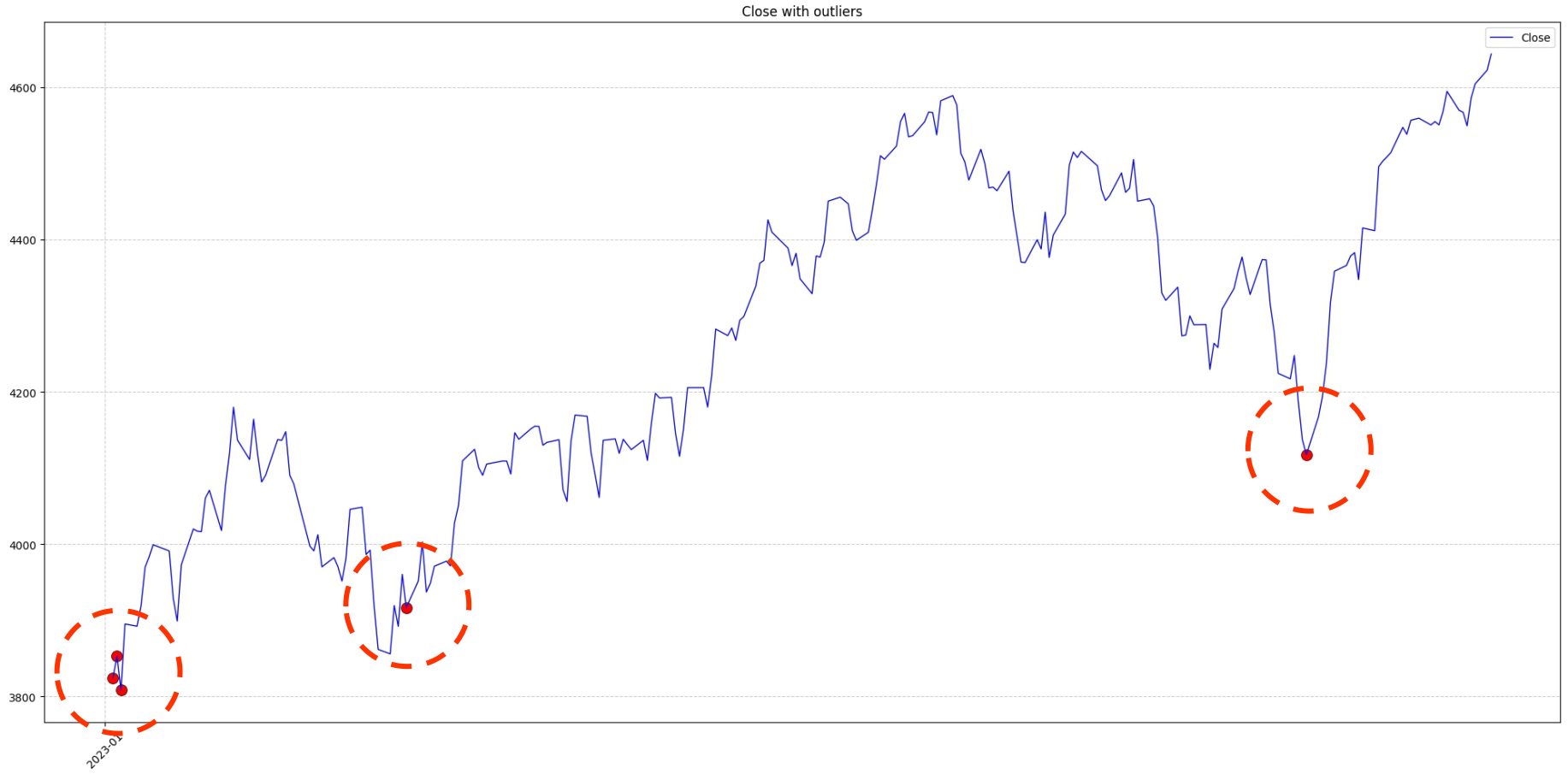
2. Exploratory Data Analysis (EDA)

Anomaly Detection

Date	**2020-03-26**	**2020-03-26**	**2020-03-26**	**2020-03-26**
z_close	True	True	False	False
iqr_close	False	False	True	False
decomp_close	False	True	False	False
arima_close	False	False	False	False
iso_multivariate	False	True	True	False
rolling_close	False	False	True	False
any_outlier	True	True	True	False
n_methods_flagged	**1**	**3**	**3**	**0**

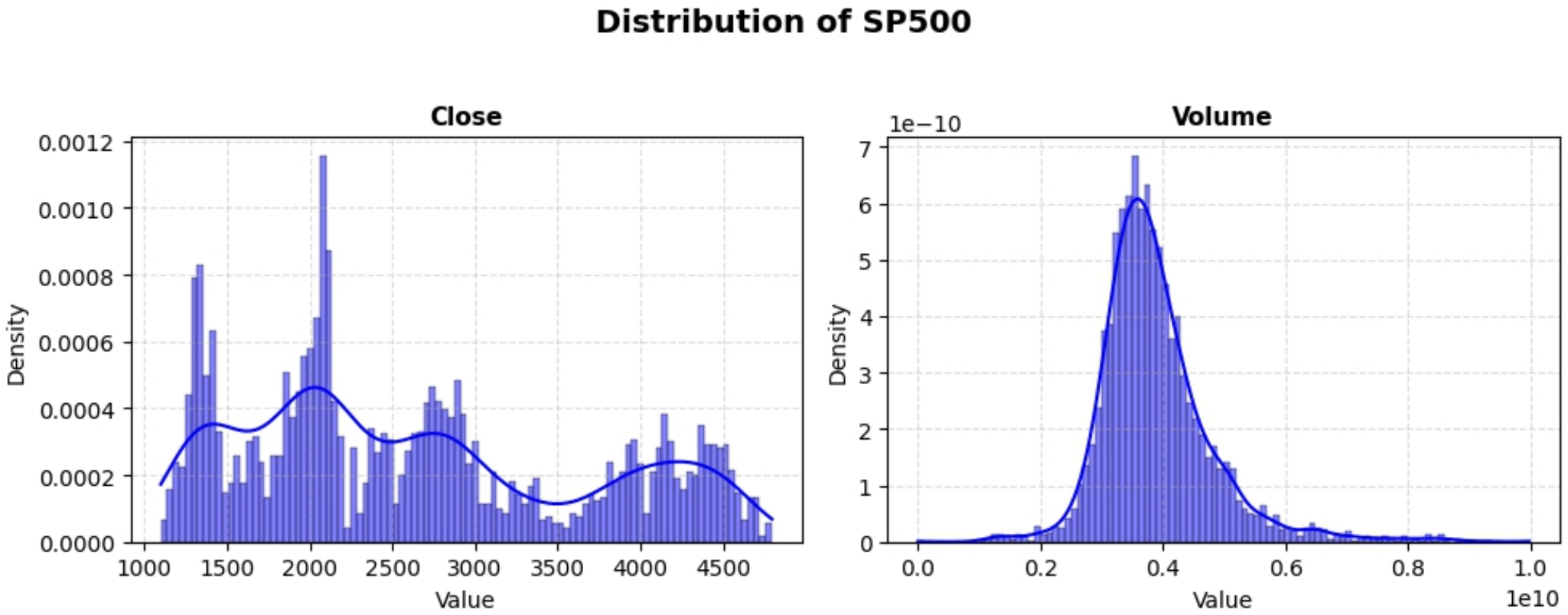
2. Exploratory Data Analysis (EDA)

Anomaly Detection



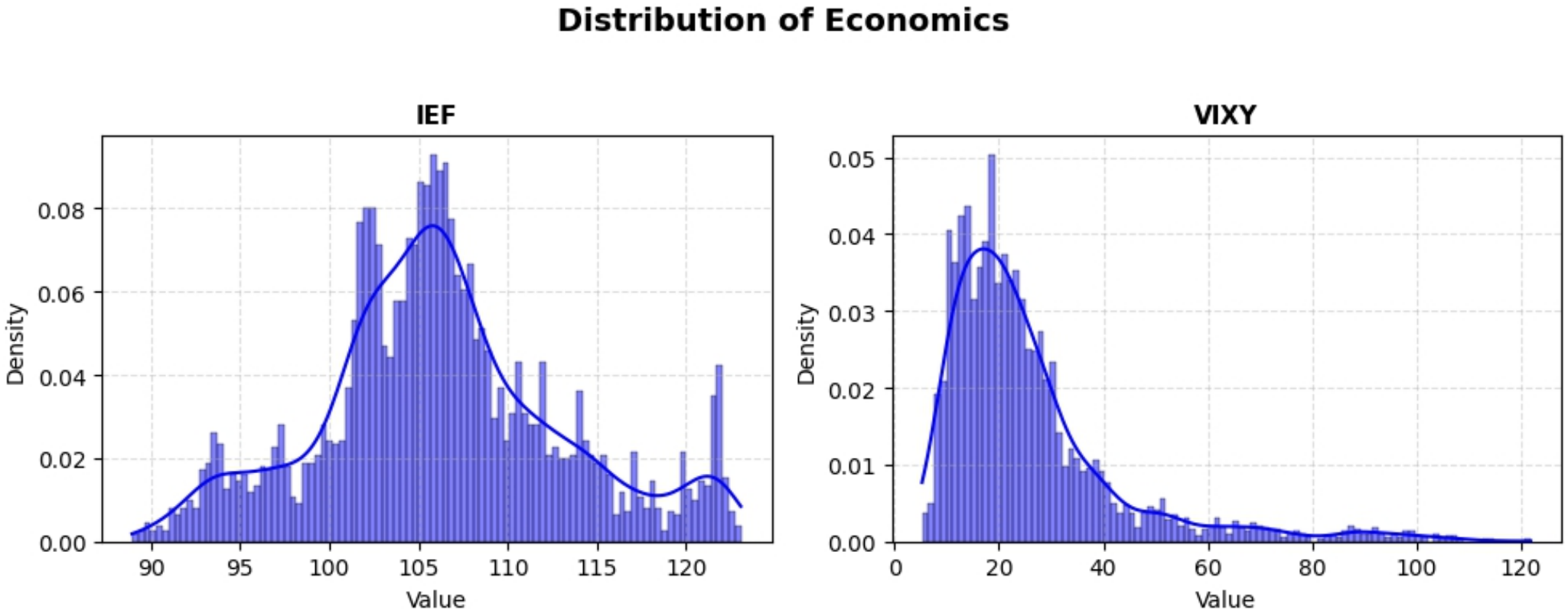
3. Proability Distribution Analysis

Close Index & Trading Volume



3. Proability Distribution Analysis

IEF & VIXY



3. Proability Distribution Analysis

Normality Test and Distribution Fit

- **Shapiro-Wilk**: Specifically designed to test normality. It evaluates how well the data's order statistics fit a normal distribution.
- **Kolmogorov-Smirnow**: A general goodness-of-fit test comparing the empirical CDF (cumulative distribution function) of the data to a reference distribution

Index	Variable	Shapiro	KS	Distribution
0	Close	2.046193e-06	3.566853e-02	Non-normal
1	Volume	2.518490e-18	1.548785e-06	Non-normal
2	IEF	7.141461e-07	5.485859e-02	Non-normal
3	VIXY	2.117785e-14	1.984101e-10	Non-normal

4. Hypothesis Testing

T-test: Is there a significant difference in trading volume between periods of low volatility and high volatility in the stock market?

- Null hypothesis: The mean trading volume during low-volatility and high-volatility periods are equal.
- Alternative hypothesis: The mean trading volumes differ between the two periods.

The dataset was divided into two groups based on the median of VIXY:

- Low-volatility period: VIXY < median(VIXY)
- High-volatility period: VIXY ≥ median(VIXY)

4. Hypothesis Testing

T-test: Result

- t-statistic: 2.325
- p-value: 0.02012

Answer: There is a statistically significant difference in trading volume between low- and high-volatility periods

4. Hypothesis Testing

ANOVA test: Is there a significant difference in the S&P 500 closing price (Close) among different economic periods?

- Null hypothesis: The mean closing prices are the same across all economic periods.
- Alternative hypothesis: At least one period has a different mean closing price.

The data was divided into three economic phases:

- Before COVID-19: 2011–2019
- During COVID-19: 2020–2021
- After COVID-19: 2022–2023

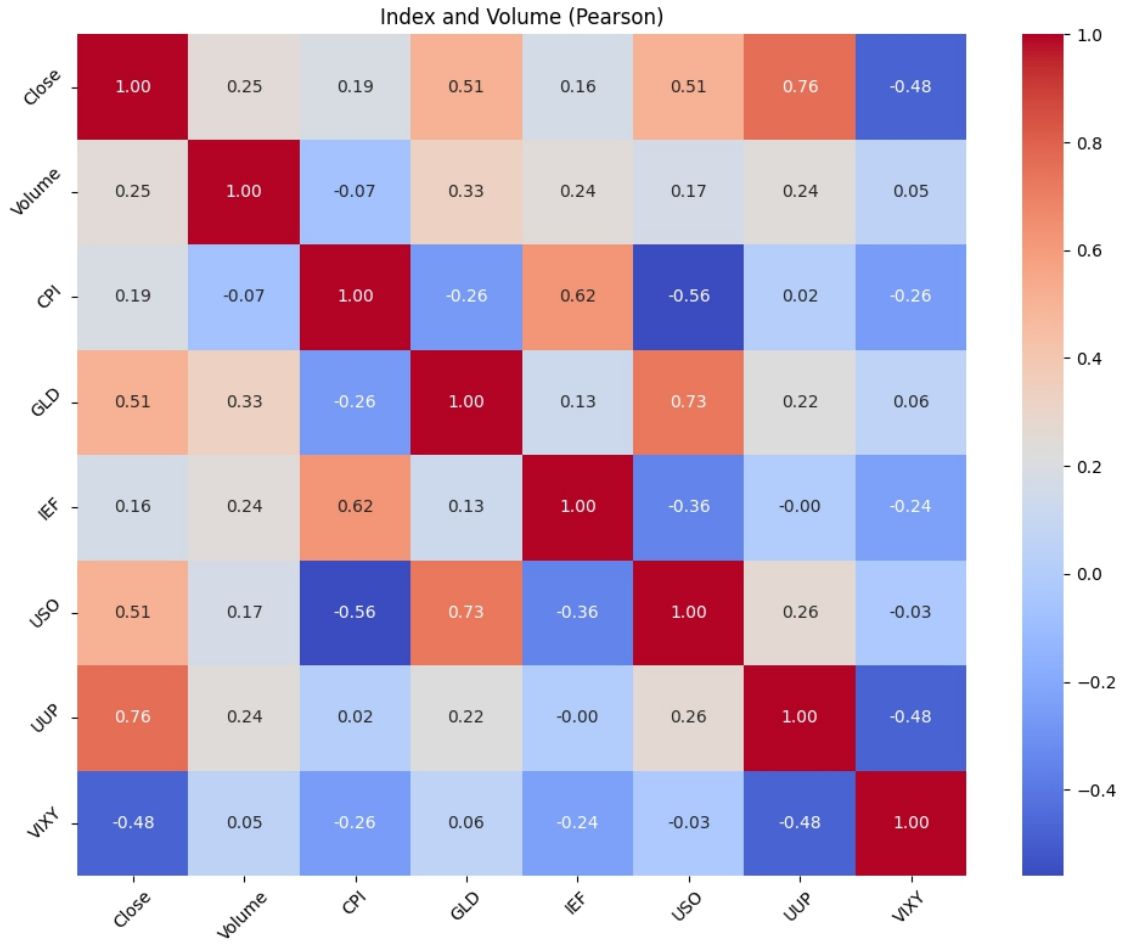
4. Hypothesis Testing

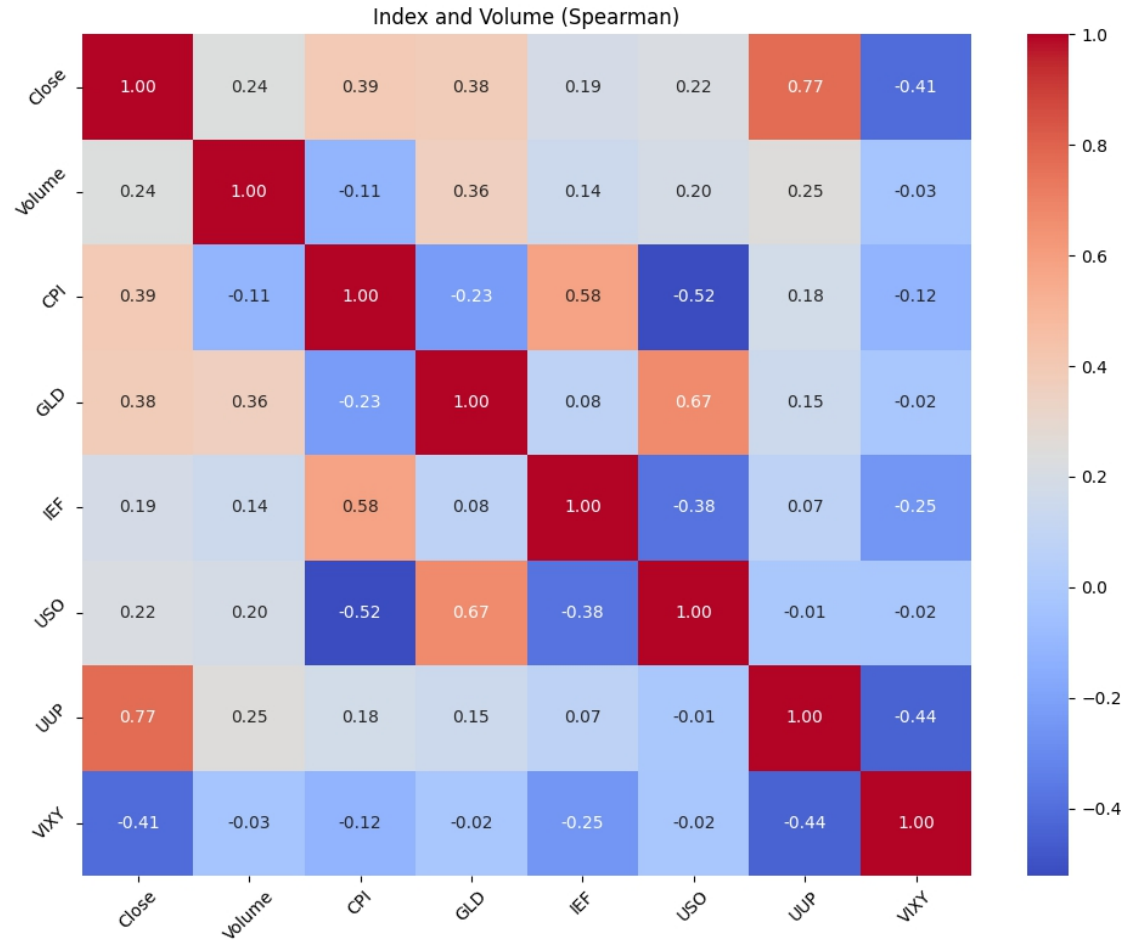
ANOVA test: Result

- f-statistic: 4588.007
- p-value: 0.0000

Answer: Therefore, there are significant differences in the mean closing price among three economic periods.

5. Correlation Analysis





6. Multiple Linear Regression

Model Implementation

- Target (y): Close
- Predictors (X): Open, High, Low, Volume, CPI, GLD, IEF, USO, UUP, VIXY, Close_lag1, Close_lag2, Close_lag3
- The model was trained using Linear Regression with a time-based train-test split (80%-20%).

6. Multiple Linear Regression

Model Evaluation

- R2: 0.9962
- RMSE: 16.9717